L Number	Hits	Search Text	DB	Time stamp
1	8396	EXCEl!	USPAT; US-PGPUB; EPO; JPO; DERWENT; IBM_TDB	2003/05/01 12:53
2	19318	HTML!	USPAT; US-PGPUB; EPO; JPO; DERWENT; IBM_TDB	2003/05/01 12:53
3	7521	HTML! NEAR2 (format$4 OR tag$2 OR file! OR files!)	USPAT; US-PGPUB; EPO; JPO; DERWENT; IBM_TDB	2003/05/01 12:55
4	86	EXCEl! SAME (HTML! NEAR2 (format$4 OR tag$2 OR file! OR files!))	USPAT; US-PGPUB; EPO; JPO; DERWENT; IBM_TDB	2003/05/01 12:55
5	36	(EXCEl! SAME (HTML! NEAR2 (format$4 OR tag$2 OR file! OR files!))) NOT @AD>19991230	USPAT; US-PGPUB; EPO; JPO; DERWENT; IBM_TDB	2003/05/01 13:02
6	1088	(HTML! NEAR2 (format$4 OR tag$2 OR file! OR files!)) NEAR5 (creat$5 OR convert$4)	USPAT; US-PGPUB; EPO; JPO; DERWENT; IBM_TDB	2003/05/01 13:00
7	14	EXCEl! SAME ((HTML! NEAR2 (format$4 OR tag$2 OR file! OR files!)) NEAR5 (creat$5 OR convert$4))	USPAT; US-PGPUB; EPO; JPO; DERWENT; IBM_TDB	2003/05/01 13:00
8	44	EXCEl! AND ((HTML! NEAR2 (format$4 OR tag$2 OR file! OR files!)) NEAR5 (creat$5 OR convert$4)) NOT @AD>19991230	USPAT; US-PGPUB; EPO; JPO; DERWENT; IBM_TDB	2003/05/01 13:04
9	42	(EXCEl! AND ((HTML! NEAR2 (format$4 OR tag$2 OR file! OR files!)) NEAR5 (creat$5 OR convert$4)) NOT @AD>19991230) NOT (EXCEl! SAME ((HTML! NEAR2 (format$4 OR tag$2 OR file! OR files!)) NEAR5 (creat$5 OR convert$4)))	USPAT; US-PGPUB; EPO; JPO; DERWENT; IBM_TDB	2003/05/01 13:04
10	3846	"Internet Explorer"	USPAT; US-PGPUB; EPO; JPO; DERWENT; IBM_TDB	2003/05/01 13:04
11	59	EXCEl! SAME "Internet Explorer"	USPAT; US-PGPUB; EPO; JPO; DERWENT; IBM_TDB	2003/05/01 13:04
12	24	(EXCEl! SAME "Internet Explorer") NOT @AD>19991230	USPAT; US-PGPUB; EPO; JPO; DERWENT; IBM_TDB	2003/05/01 13:07
13	81	EXCEl! NEAR20 (browser$2 OR web! OR hypertext! OR hyper-text! OR html!) NEAR20 format$4	USPAT; US-PGPUB; EPO; JPO; DERWENT; IBM_TDB	2003/05/01 13:07
14	32	(EXCEl! NEAR20 (browser$2 OR web! OR hypertext! OR hyper-text! OR html!) NEAR20 format$4) NOT @AD>19991230	USPAT; US-PGPUB; EPO; JPO; DERWENT; IBM_TDB	2003/05/01 13:07

US-PAT-NO: 6549898

DOCUMENT-IDENTIFIER: US 6549898 B1

TITLE: Method of and an apparatus for retrieving and delivering documents and a recording media on which a program for retrieving and delivering documents are stored

---------- KWIC ---------

In this technique, a keyword (called "feature character string" in prior art 2) is extracted from a text (to be referred to as a seed text) exemplified as a **retrieval** condition to calculate **similarity of the seed document with respect to registered documents**.

In accordance with prior art 2, feature character strings (keywords) are extracted from a seed **document and then similarity of each registered document** to the seed document is calculated using the keywords. The user specifies a document containing information desired by the user such that the user then refers to results of **retrieval** in the descending order of similarity to obtain texts containing necessary information from the text database.

DOCUMENT-IDENTIFIER: US 20010047351 A1

TITLE: Document information search apparatus and method and recording medium storing document information search program therein

---------- KWIC ---------

An apparatus for searching document information by a server in response to a search request from a client and responding. In the case where a document file is designated as a search condition by a search condition designating unit of the client, contents of a **designated file** are transmitted via a network. A document search unit of a search machine provided on the server side forms a keyword from the file contents transmitted from the search condition designating unit and searches similar documents from an index (a train of important words extracted from search target documents) in a search database.

Abe, Seiichiro

[0008] According to the invention, there is provided a document information search apparatus which searches document information and responds on a search side such as a server or the like on the basis of a search request sent from a client or the like via a network, wherein: a search condition designating unit which, when a file is designated as a search condition, transmits contents of the **designated file** through the network is provided for a requesting source such as a client or the like; and a search machine which forms a keyword from the file contents transmitted from the search condition designating unit and searches similar documents is provided on a search side such as a server or the like. Therefore, if the user wants to obtain documents including interesting contents by Email, Internet, or the like and search documents of contents similar to such a document or the like, a file which was uploaded by a designation of a document is designated as a search condition, thereby enabling the documents having similar contents to be searched. Therefore, any document which is not registered in a database can be freely designated as a search condition, a troublesome input of keywords based on the document contents becomes unnecessary, and the similar documents can be easily and promptly searched.

[0009] The search condition designating unit on the search requesting source transmits a **head file** portion of the **designated file** contents. Since many important keywords necessary for document search exist usually in a head portion of a document, only the head portion of the file contents, for example, the head portion of 1 kB is transmitted as a search condition. Since the document files which are used for the search condition have various sizes, by deciding a capacity of the file which is transmitted as a search condition, a communication load and the processes on the search side are reduced. The search condition designating unit includes an HTML file and an Excel file as files which are designated as a search condition. Even in the other file formats, the files include a file of an arbitrary file format so long as it is a file from which a text document can be extracted. A database in which index information describing a list of important words extracted from the search target documents has been stored every document is provided for the search machine on the server side. The document search unit of the search machine comprises: a text extraction processing unit which extracts a text document from the file contents received in response to the search request; a morpheme analyzing unit which extracts nouns by a morpheme analysis of the text document; a keyword forming unit which extracts important words from the nouns and forms a keyword in which the important words are coupled by OR; and a search executing unit which searches similar documents by searching the search database by the keyword and notifies the client of a search result. The

keyword forming unit counts the number (H) of times of appearance showing in which document in the index of each search document stored in the search database each noun appears and selects a predetermined number of upper words each having the number (H) of times of appearance in a predetermined range, thereby forming the keyword. When the number of documents in the index is assumed to be N, the keyword forming unit selects upper ten words each having the number of times of appearance in a range in which the number (H) of times of appearance is equal to, for example,

[0012] The invention provides a document information search method of searching document information and responding on a search machine side such as a server or the like on the basis of a search request which is transmitted from a search requesting source such as a client or the like via a network, comprising the steps of: storing index information describing a list of important words extracted from search target documents into a search database of the server every document; when a document file is designated as a search condition, transmitting contents of a **designated file** to a search side via the network together with the search request; and on the search side, extracting a text document from the received file contents in response to the search request, extracting nouns by a morpheme analysis of the text document, extracting important words from the nouns, forming a keyword in which the important words are coupled by OR, searching similar documents by searching a search database by the keyword, and notifying the client of a search result. The details of the document information search method are fundamentally the same as those of the apparatus construction.

[0028] FIG. 2 is a block diagram of a functional construction in the search system of FIG. 1. First, a search condition designating unit 26 is provided for the WWW browser 16 serving as a user side. The search condition designating unit 26 of the invention directly designates a document file, as a search condition, obtained by the user as a search condition via the Internet, Email, or the like, and transmits the contents of the **designated file** to a document search unit 30 of the search machine 20 via the WWW server 18 through the Internet/Intranet 14. Besides the search condition of the file designation which is newly provided in the invention, the search condition designating unit 26 can also designate the following search conditions.

[0035] FIG. 3 shows the details of a functional construction of the document search unit 30 of the invention provided for the search machine 20 in FIG. 2. A search designation file storing unit 34, a text extraction processing unit 36, a morpheme analyzing unit 38, a keyword forming unit 40, and a search executing unit 42 are provided for the document search unit 30. An index 52 comprising a set of important words, the document name, storing location, and the like of each of the search target documents 25 in the document database 24 formed by the search database forming unit 28 in FIG. 2 has been stored in the search database 22. The file contents transmitted by the file designation of the search condition designating unit 26 in the WWW browser 16 in FIG. 2 are stored in the search designation file storing unit 34 in the document search unit 30. When the file contents are transferred from the WWW browser 16 side, a **head file** portion of the document file designated as a search condition, for example, 1 kB of the head portion is extracted and transmitted together with the search request to the WWW server 18 side. A capacity of the file which is transmitted as a search condition is set to a fixed capacity, for example, 1 kB as mentioned above, thereby setting a transfer load of the document contents to the search machine 20 side to be constant irrespective of a size of document file designated as a search condition. The searching process by the document search unit 30 in the search machine 20 is stabilized and a high processing speed is realized. The text extraction processing unit 36 extracts a text document from the file contents designated as a search condition stored in the search designation file storing unit 34. As a format of the document file which is designated as a search condition in the WWW browser 16, there are various file formats such as text file of Email, HTML file in the Internet, further, Excel file of an aggregate list, and the like. Therefore, to enable a search function to be presented with respect to a difference of the file

formats, only the text document is extracted from the document files of various formats by the text extraction processing unit 36 and used as a search condition. The morpheme analyzing unit 38 subsequently provided extracts nouns included in the extracted text document by using a morpheme analysis. The nouns in the document contents extracted by the morpheme analyzing unit 38 are sent to the keyword forming unit 40. The keyword forming unit 40 extracts important nouns in order to form a keyword. As for the extraction of the important words in the keyword forming unit 40, first, the number (H) of times of appearance showing in which documents in the number (N) of documents registered in the index 52 in the search database 22 each noun appears is counted. When the number (H) of times of appearance of document in the index 52 is obtained, words in which the number (H) of times of appearance lies within a predetermined range, for example,

[0036] are selected. Upper ten words in which the number (H) of times of appearance is large among the words selected as mentioned above are selected to form the keyword. A query expression in which the 10 selected important words are coupled by OR is formed and provided to the search executing unit 42. On the basis of the query expression derived from the keyword forming unit 40, the search executing unit 42 performs a search collation with the index 52 in the search database 22, extracts an index which satisfies a predetermined similarity as a search result, and transmits the search result to the WWW browser 16 side by the WWW server 18, thereby enabling the user to refer to the search result in a form of a document list. Further, the document search unit 30 can also perform a document search using property information of the file designated as a search condition stored in the search designation file storing unit 34. For this purpose, when the document file is designated as a search condition, the search condition designating unit 26 in the WWW browser 16 extracts the property information of the designated document file and transmits the property information to the search machine 20 side together with the **head file** portion, for example, 1 kB of the **head file** portion of the document designated as a search condition. In the document search unit 30 in FIG. 3, in addition to the extraction of the text document from the file contents, the extraction of the nouns by the morpheme analysis, and the formation of the keyword by the selection of the important words with respect to the nouns, for example, a date of formation, a writer, a title, and the like are extracted from the property information added to the file contents stored in the search designation file storing unit 34. The property information is included in the keyword by the keyword forming unit 40. The index 52 in the search database 22 is searched by the search executing unit 42.

[0038] FIG. 5 is a flowchart for a browser process for performing the designation of the search condition and the display of the search result by the WWW browser 16 in FIG. 2. When the user opens the search function of the WWW browser 16, a search picture plane is displayed in step S1. When the search picture plane is displayed, a designating operation of the search condition in which the document file has been designated is performed in step S2. Subsequently, in step S3, whether the search has been activated or not is discriminated. When the search activation is determined, whether the search is a file designating search or not is discriminated in step S4. If YES, step S5 follows and the file designated by the user is read out. In step S6, 1 kB of the head in the **designated file** is transmitted to the server together with the search requesting message. If the search is not the file designating search, a search requesting message corresponding to the other search, for example, a keyword search is transmitted to the server in step S7. When the head portion in the **designated file** is transmitted to the server in step S6, the apparatus waits for reception of the search result in step S8. When the search result is received from the server in step S8, step S9 follows and the user executes a display operating process of the search result and looks at the search contents. Such processes in steps S1 to S9 are repeated until a search end instruction for closing the search picture plane is issued in step S10.

1. A document information search apparatus for searching document information on the basis of a search request transmitted through a network and

responding, wherein: a search condition designating unit which designates a file as a search condition and transmits contents of said **designated file** via the network is provided for a search requesting source; and a document search unit which forms a keyword from the file contents transmitted from said search condition designating unit and searches similar documents from a database is provided on a search side.

2. An apparatus according to claim 1, wherein said search condition designating unit transmits a **head file** portion of the **designated file** contents.

13. A document information search method of searching document information on the basis of a search request transmitted via a network and responding, comprising the steps of: storing index information describing a list of important words extracted from search target documents every document into a database; in the case where a file is designated as a search condition on a search requesting source, transmitting contents of the **designated file** to a server together with the search request through the network; and on a search side, extracting a text document from the file contents received in response to the search request, extracting nouns by a morpheme analysis of said text document, extracting important words from said nouns, forming a keyword in which said important words are coupled by OR, searching similar documents by searching said database by said keyword, and responding a search result.